A Appendix: Additional Results

Appendix Figures A.1 and A.2 evaluate the selection argument for the rising educational gradient in midlife mortality. Appendix Figure A.1 plots the education-specific mortality rate (y-axis) on education percentiles (x-axis) by year for college and non-college graduates. The figure demonstrates that the education percentiles of college and non-college populations, defined here as the midpoint of each education group's percentile range, changed little, falling from the 88th to 83rd percentile for college graduates and from the 38th to the 33rd percentiles for non-college graduates. Yet mortality rates decreased monotonically for college graduates from 1992 to 2019 (from 262 to 133 deaths per 100,000). For non-college graduates, the decline slowed after 2000, and deaths per 100,000 rose after 2010, such that mortality *increases* slightly comparing 1992 with 2019 (458 versus 463 deaths per 100,000). Appendix Figure A.2 uses the bounds-based selection correction of Novosad, Rafkin, and Asher (2022) to provide additional evidence that mortality patterns were not driven by selection. In the raw data, the widening educational mortality gradient is driven by a relatively stable non-college rate and a declining college rate. The green areas in each panel of the figure show that if we define the high-education group to be persons in the top 30 percent of educational attainment, rather than college graduates, we also see declines in mortality over time. The blue and red areas provide bounds for persons in the bottom 10 percent of the educational distribution and the 10th–70th percentiles. These areas are essentially stable, as is the national non-college mortality rate in the unadjusted data.

Appendix Figure A.3 shows that the mortality gap increase for Black and Hispanic populations is similar to the increase for the entire population. In addition, increases in the Black and Hispanic gaps arise from declining college mortality rates combined with more-stable non-college rates, which is also the pattern in the overall population.

Appendix Figure A.4 shows that mortality patterns for a largely non-immigrant population (Blacks and Whites) are nearly identical to the pattern for the entire population. The implication is that full-sample results are not likely to be biased by states with large numbers of immigrants, who tend to have lower mortality rates.

Appendix Figure A.5 displays a map of county codes from the 2013 Urban-Rural Classification Scheme from the National Center for Health Statistics.

Appendix Tables A.1 and A.2 present county-level mortality-change regressions for non-college and college populations excluding the rural/urban controls. (Regressions with these controls appear as Tables 1 and 2.) This sensitivity analysis is motivated by the potential for measurement error. If a right-hand side variable is measured with error, then some of its true effect may be absorbed within the urban-rural controls.

Appendix Table A.6 provides the state-level analysis of mortality gaps that is referenced in footnote 21. The horizontal axis of each chart depicts a prediction for the change in a state's mortality-education gap between 1992 and 2019. This prediction is constructed by subtracting the mortality-change prediction from a college regression model in Table 3 from that from a non-college model. The vertical axes depict actual mortality-gap changes. The top chart uses models that contain only smoking variables (columns 2 and 6 of Table 3). The bottom chart uses predictions from models with both the smoking and income variables (columns 3 and 7). A comparison of the two charts shows that most of the information needed to predict the change in a state's mortality gap is contained in the smoking variables.

Appendix Table A.3 and Appendix Figures A.7 and A.8 replicate the state-level analysis of Table 3, Figure 11, and Appendix Figure A.6 using regression models that also include 1992 education-specific mortality rates as right-hand-side variables. The inclusion of the lagged mortality rate in the college regressions significantly improves the mortality-change predictions for those populations, due to the strong beta convergence in college mortality. But the main lesson of the baseline state-level analysis survives, as smoking behavior alone (particularly non-college smoking) does a good job of predicting how state-level mortality evolves between 1992 and 2019.

Appendix Figure A.9 replicates the analysis of potential convergence in education-specific state-level smoking rates in Figure 13 using data from the CPS-TUS. (Figure 13 uses data from BRFSS.)

Appendix Figures A.10 and A.11 present daily 2014 smoking rates for the United States and five European countries. The first of these figures corresponds to all education levels and the second is disaggregated by college/non-college. The number at the top of each bar is the total daily smoking rate for the given category of persons. This number is equal to the combined height of the red and blue sections of the bar. The lower number corresponds to the share of persons who are daily smokers that smoke more than 20 cigarettes per day. The number corresponds to the blue section of the bar.

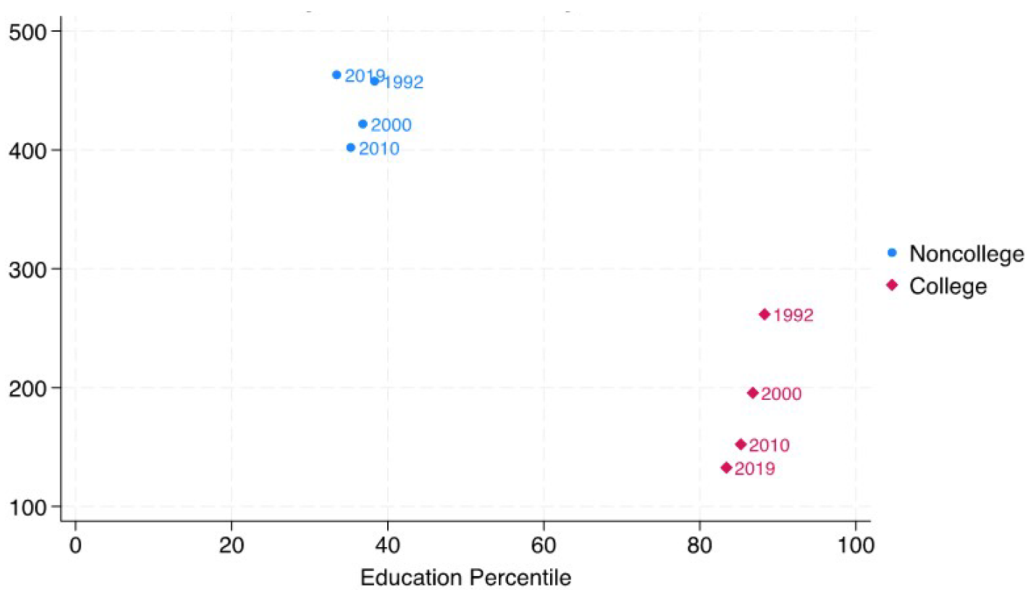

Figure A.1. Midlife Mortality Rates by Education and Education Percentile. Education percentile is defined as the midpoint of each education group's percentile range. Mortality rates are age-adjusted deaths per 100,000 population among adults aged 25–64.

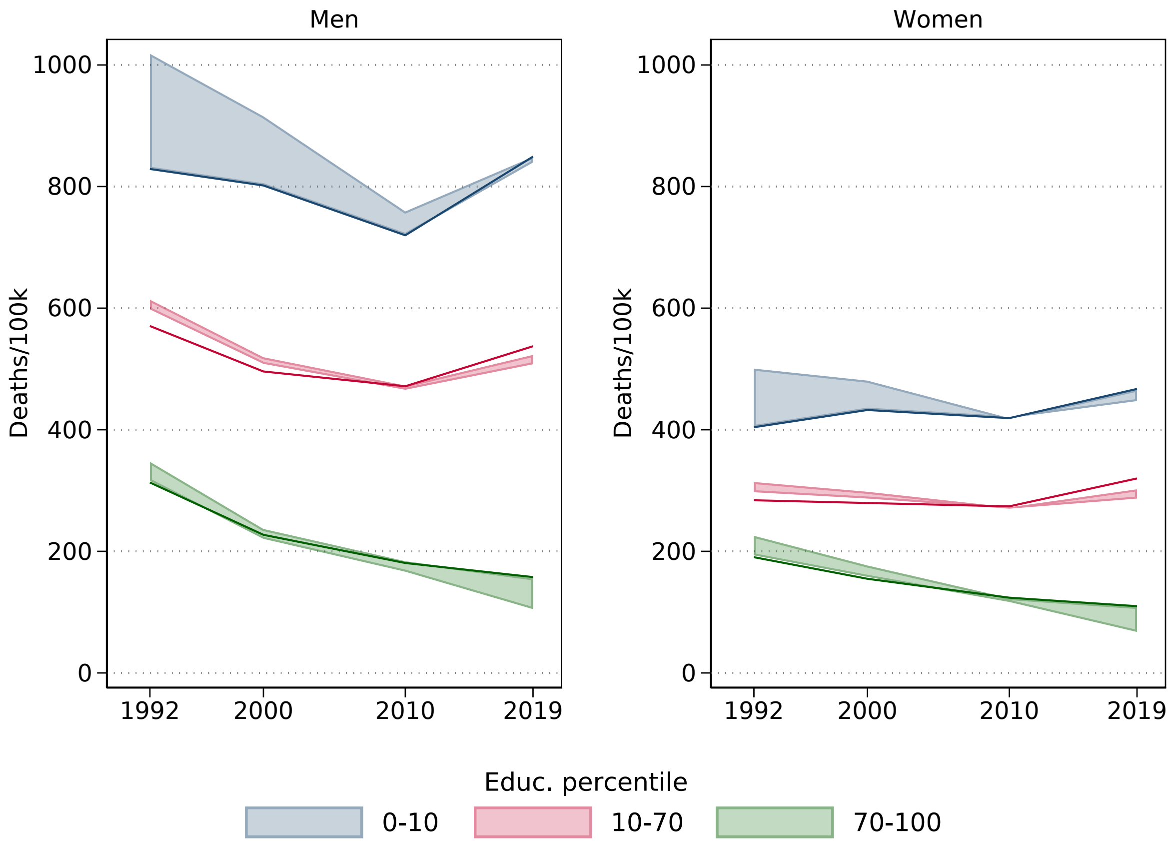

Figure A.2. Midlife Mortality by Education and Year with Novosad et al. (2022) Selection Correction. Notes: The selection model in Novosad, Rafkin, and Asher (2022) develops bounds on the education-mortality gradient that are anchored on the 10th and 70th percentiles. Assuming that health is monotonically related to education percentiles, the authors are able to estimate upper and lower bounds on the percentile estimates. These bounds are widely spaced when the actual rate of (say) high school dropouts is different from the 10th education percentile, as it is for men in 1992.

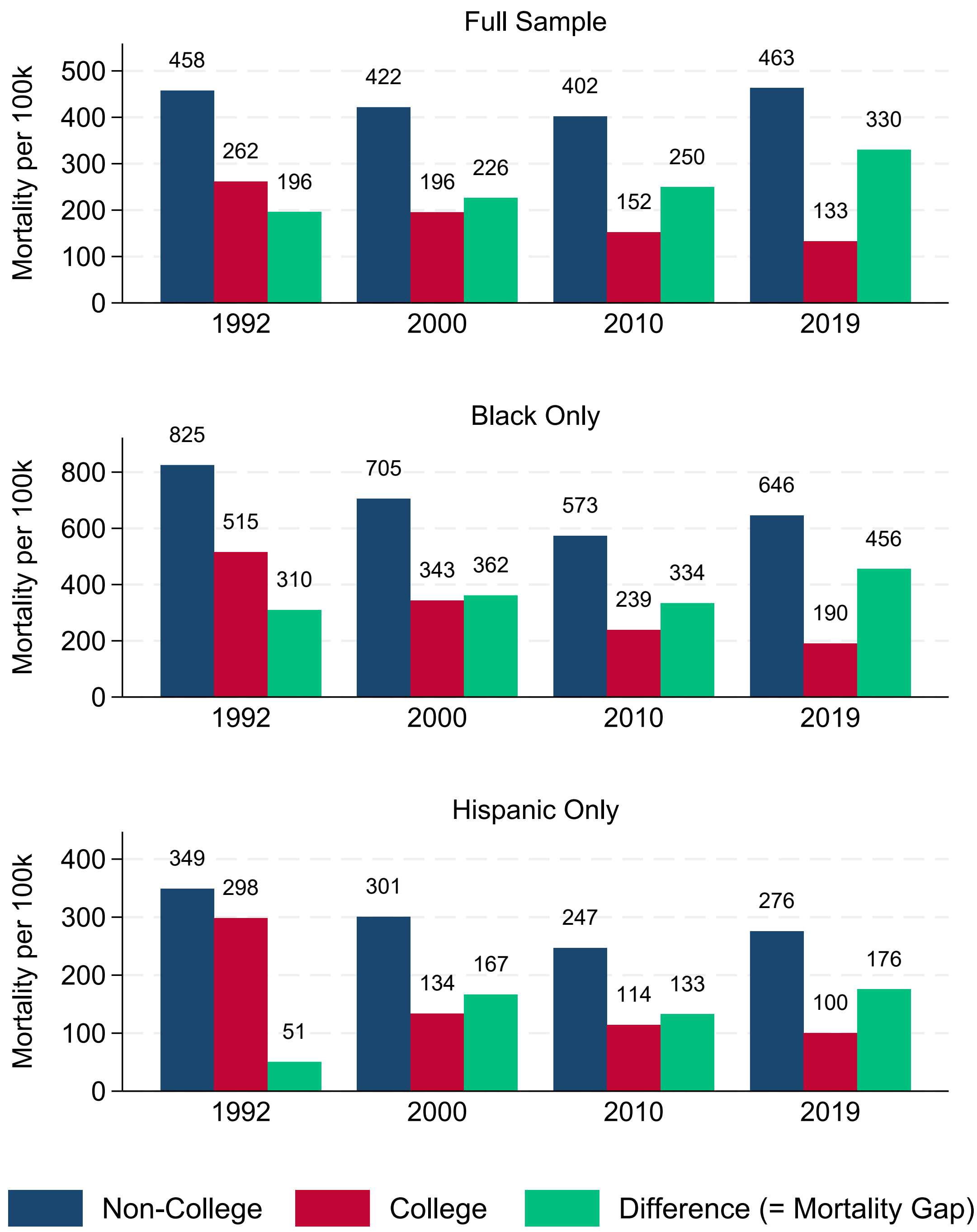

Figure A.3. Midlife Mortality Rates by Education and Year for the Full Sample and non-Hispanic Black and Hispanic Populations of Any Race Separately.

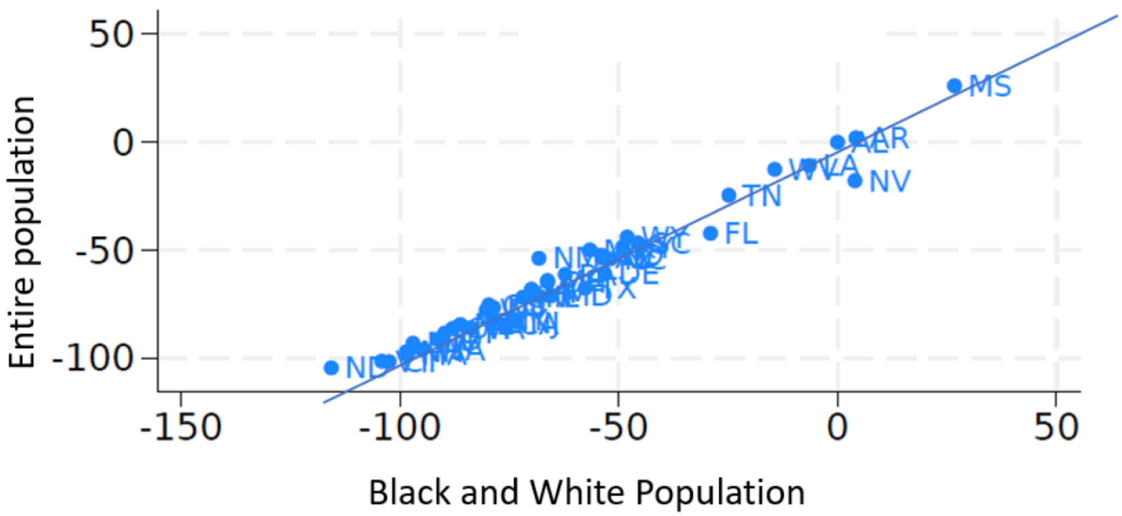

Figure A.4. State-Level Mortality for Non-Hispanic Black and Non-Hispanic White Populations Compared to the Full Population. Notes: The mortality estimates are centered on Alaska.

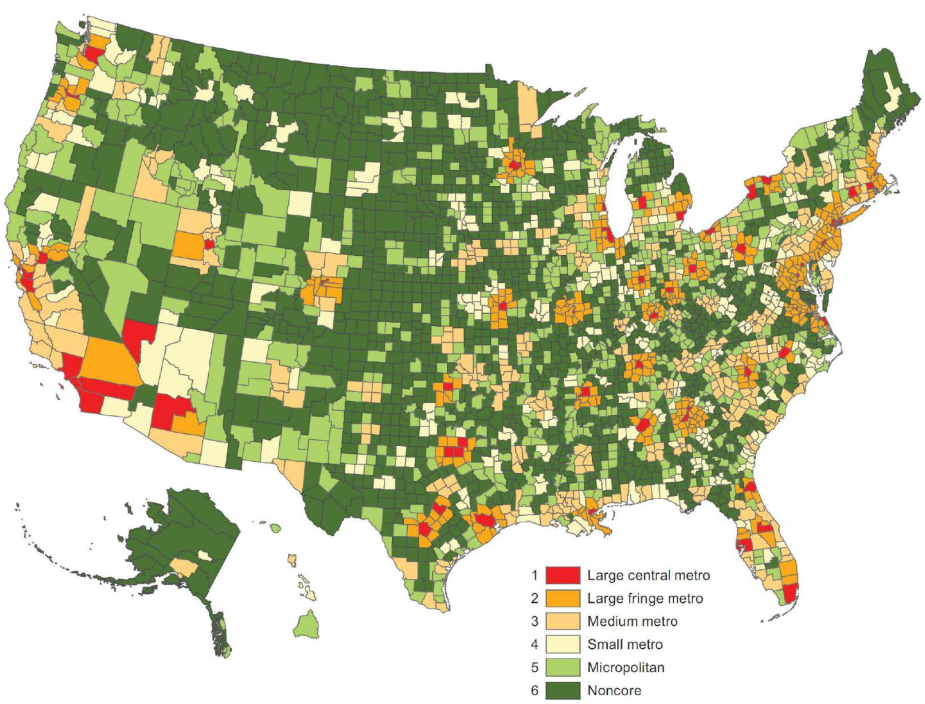

Figure A.5. County Codes from the 2013 Urban-Rural Classification Scheme from the National Center for Health Statistics.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1992 Mortality Rate	–26.61***	–26.15***	–36.28***	–37.30***	–34.38***	–26.24***	–25.25***	–26.01***	–32.77***	–43.57***	–41.60***
	(6.174)	(6.090)	(5.304)	(5.758)	(5.157)	(5.990)	(5.828)	(6.205)	(5.572)	(4.108)	(5.435)
Avg. 1-yr Outmig Rate (2016-20)		9.634								0.462	–2.432
		(8.559)								(3.885)	(3.939)
Smoking Rate			63.10***							51.41***	42.57***
			(4.979)							(4.901)	(7.132)
Obesity Rate				34.65***						8.954*	10.89***
				(4.930)						(3.685)	(2.924)
Ln 1992 Per Capita Income					–38.53***					–27.85***	–21.67**
					(3.661)					(6.242)	(6.347)
1993-2019 Chg Ln Per Cap Income					–27.31***					–12.88***	–14.29***
					(3.750)					(2.581)	(2.801)
1992 Emp-Pop Ratio						–14.81***				6.950*	3.206
						(3.647)				(3.321)	(3.646)
1993-2019 Chg Emp-Pop Ratio						–31.35***				5.879	3.826
						(7.370)				(4.886)	(4.777)
1992 Manufacturing Empl Share							24.35***			2.856	–0.420
							(3.678)			(3.616)	(3.562)
China Shock								9.541^{+}		0.258	–2.049
								(4.784)		(2.726)	(2.432)
State Policy-Liberalism Index									–35.40**	–5.071	
									(11.86)	(5.047)	
Constant	–16.08	–16.08	–16.08**	–16.08	–16.08	–16.08	–16.08	–16.08	–16.08	–16.08***	28.91***
	(15.85)	(15.21)	(5.455)	(14.80)	(11.66)	(13.77)	(14.15)	(15.69)	(11.61)	(4.491)	(4.879)
N	2081	2081	2081	2081	2081	2081	2081	2081	2081	2081	2081
R-sq	0.0590	0.0668	0.383	0.150	0.273	0.163	0.108	0.0666	0.160	0.477	0.553
State FEs?	N	N	N	N	N	N	N	N	N	N	Y

Table A.1. County-Level Mortality-Change Regressions, with Standardized Coefficients: Non-College. + 0.10 * 0.05 ** 0.01 *** 0.001. The dependent variable is 1993–2019 mortality-rate change for the county's non-college population. Standard errors are clustered by state and the data are weighted by education-specific average population in 1993 and 2019. The average 1-year out-migration rate is a 5-year average from the 2016–2020 American Community Survey. The China shock variable is from Autor, Dorn, and Hanson (2013), and the state policy-liberalism index is from Caughey and Warshaw (2016). Because the regressors are standardized, coefficients denote effects of 1-standard-deviation increases in the regressors.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
1992 Mortality Rate	–33.78***	–34.00***	–35.46***	–34.88***	–33.46***	–31.41***	–32.59***	–33.60***	–34.27***	–31.01***	–26.80***
	(8.037)	(8.060)	(8.641)	(7.149)	(6.798)	(6.094)	(7.833)	(8.027)	(7.916)	(6.123)	(6.486)
Avg. 1-yr Outmig Rate (2016-20)		–2.743								1.986	–0.681
		(2.749)								(3.320)	(4.091)
Smoking Rate			6.268							–0.526	–1.597
			(3.897)							(1.981)	(2.465)
Obesity Rate				16.80***						10.68***	7.266**
				(3.901)						(2.360)	(2.538)
Ln 1992 Per Capita Income					–5.086*					5.930***	5.382*
					(2.367)					(1.659)	(2.050)
1993-2019 Chg Ln Per Cap Income					–17.02***					–13.94***	–12.31***
					(1.833)					(1.556)	(1.662)
1992 Emp-Pop Ratio						–9.081**				–5.416+	–8.014**
						(2.833)				(2.935)	(2.767)
1993-2019 Chg Emp-Pop Ratio						–11.73**				–1.097	–2.802
						(3.868)				(2.972)	(3.363)
1992 Manufacturing Empl Share							10.42***			6.235**	4.899*
							(1.600)			(1.893)	(1.868)
China Shock								3.508*		2.096	3.955+
								(1.734)		(1.922)	(1.972)
State Policy-Liberalism Index									–6.349	–0.166	
									(5.937)	(4.267)	
Constant	–123.7***	–123.7***	–123.7***	–123.7***	–123.7***	–123.7***	–123.7***	–123.7***	–123.7***	–123.7***	–106.6***
	(5.356)	(5.509)	(4.759)	(3.836)	(3.372)	(4.821)	(5.079)	(5.427)	(4.526)	(2.922)	(3.180)
N	2081	2081	2081	2081	2081	2081	2081	2081	2081	2081	2081
R-sq	0.184	0.185	0.189	0.229	0.241	0.219	0.201	0.186	0.190	0.265	0.303
State FEs?	N	N	N	N	N	N	N	N	N	N	Y

Table A.2. County-Level Mortality-Change Regressions, with Standardized Coefficients: College. + 0.10 * 0.05 ** 0.01 *** 0.001. The dependent variable is 1993–2019 mortality-rate change for the county's college population. Standard errors are clustered by state and the data are weighted by education-specific average population in 1993 and 2019. See the notes to Table A.1 for a description of the average 1-year out-migration rate and the sources of the child economic-connectedness and state policy-liberalism regressors. Because the regressors are standardized, coefficients denote effects of 1-standard-deviation increases in the regressors.

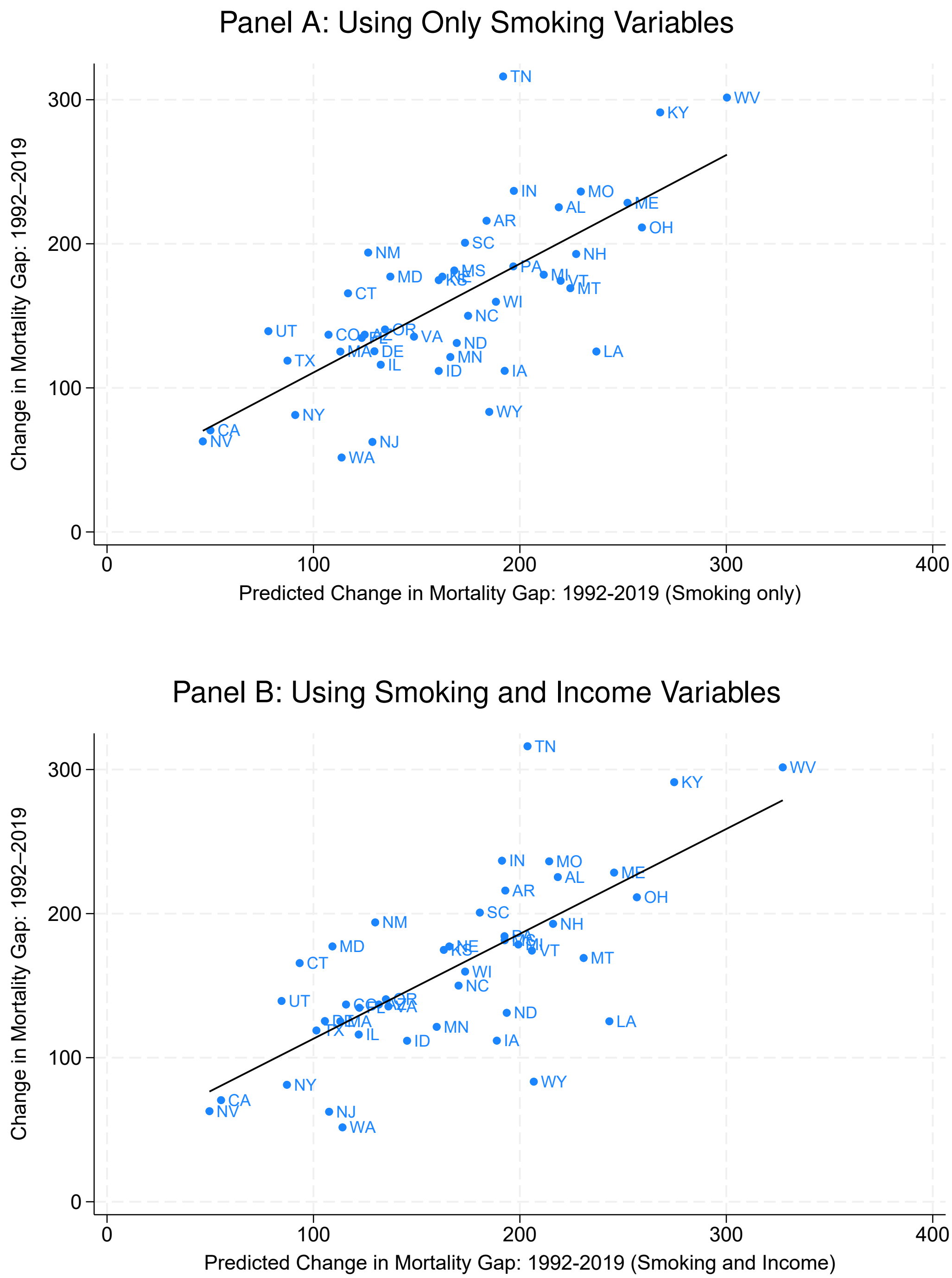

Figure A.6. Predicted and Actual Values for 1992–2019 State-Level Changes in Mortality-Rate Gaps. Predicted values for mortality-gap changes in Panel A are the differences between non-college mortality-change predictions generated by column 2 of Table 3 and college mortality-change predictions generated by column 6. The predicted values for Panel B are the differences between non-college predicted values generated from column 3 and college predicted values generated from column 7.

	(1) Non-Coll	(2) Non-Coll	(3) Non-Coll	(4) Non-Coll	(5) College	(6) College	(7) College	(8) College
Ln 1992 State Per-Capita Income	–340.2***		–151.2*	–158.9*	–88.24***		–71.78***	–85.63***
	(70.81)		(68.30)	(74.09)	(14.84)		(16.71)	(22.02)
1992-2019 Chg. in Ln State Income	–674.3***		–109.3	–40.09	–73.74*		–30.84	–42.60
	(165.0)		(127.6)	(131.0)	(31.24)		(37.29)	(42.67)
1992 Non-College Smoking Rate		1189.7***	981.1***	1070.5***				
		(168.0)	(220.4)	(257.3)				
1992–2019 Chg in Non-College Smoking Rate		1273.8***	1012.6***	1034.5***				
		(123.9)	(119.7)	(124.2)				
1992 Non-College Obesity Rate				–402.0				
				(278.7)				
1992–2019 Chg in Non-College Obesity Rate				171.5				
				(209.5)				
1992 College Smoking Rate						883.9***	414.0*	395.2
						(196.9)	(196.4)	(231.4)
1992–2019 Chg in College Smoking Rate						926.9***	440.9*	374.2
						(186.9)	(177.3)	(227.1)
1992 College Obesity Rate								–124.2
								(83.02)
1992–2019 Chg in College Obesity Rate								–3.346
								(69.36)
1992 Non-College Mortality Rate	–0.531***	–0.402*	–0.435**	–0.414**				
	(0.134)	(0.159)	(0.147)	(0.141)				
1992 College Mortality Rate					–0.863***	–0.887***	–0.880***	–0.878***
					(0.0479)	(0.0933)	(0.0507)	(0.0501)
Constant	4118.2***	0.632	1671.9*	1716.4*	1050.8***	70.28**	849.0***	1007.7***
	(767.8)	(78.59)	(782.3)	(844.9)	(152.8)	(23.42)	(183.4)	(254.4)
N	44	44	44	44	44	44	44	44
R-sq	0.628	0.787	0.826	0.841	0.916	0.880	0.925	0.928

Table A.3. State-Level Mortality-Change Regressions Including 1992 Mortality Rate. The dependent variable for each regression is the state-level mortality-rate change from 1992 to 2019 for non-college (columns 1–4) or college (columns 5–8) populations. Data are weighted by average education-specific population in 1992 and 2019 and robust standard errors are in parentheses. State-level smoking rates for 1992 and for 2019 (used for the 1992–2019 change) are calculated from BRFSS.

Figure A.7. Predicted and Actual Values for 1992–2019 State-Level Education-Specific Mortality-Rate Changes, From Models with 1992 Mortality Rates Included. Predicted values for Panel A use the regression models in columns 2 (for non-college) and 6 (for college) in Table A.3. These models use only the 1992 level and the 1992–2019 change in education-specific smoking rates. Panel B uses predictions from the models in columns 3 and 7, which add the 1992 level and 1992–2019 change in state per-capita income to the regression models.

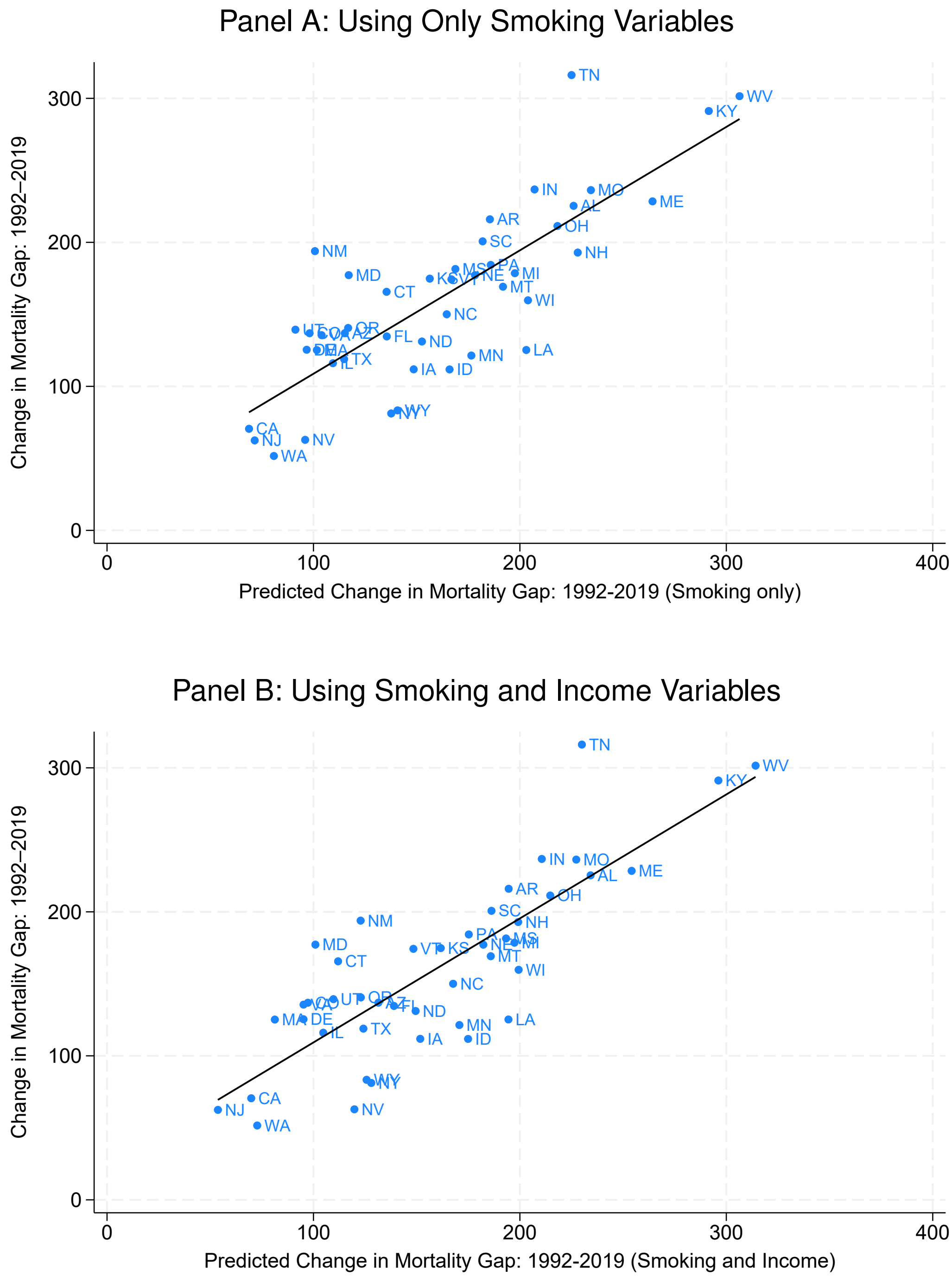

Figure A.8. Predicted and Actual Values for 1992–2019 State-Level Changes in Mortality-Rate Gaps, from Models with 1992 Mortality Rates Included. Predicted values for mortality-gap changes in Panel A are the differences between non-college mortality-change predictions generated by column 2 of Table A.3 and college mortality-change predictions generated by column 6. The predicted values for Panel B are the differences between non-college predicted values generated from column 3 and college predicted values generated from column 7.

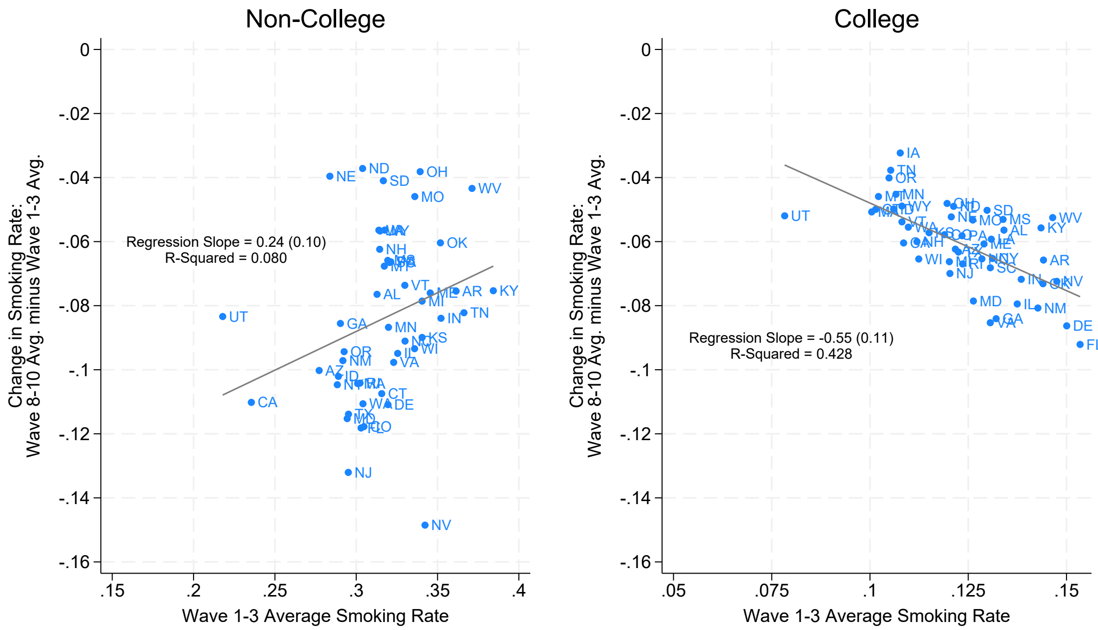

Figure A.9. State-Level Changes in Current-Smoking Rate at Midlife and Initial Current-Smoking Rate by Education, using CPS-TUS data. *Notes:* Data are age-adjusted and regressions are unweighted. *Source:* Current Population Survey Tobacco-Use Supplements Waves 1 (1992–1993), 2 (1995–1996), 3 (1998–1999), 8 (2010–2011), 9 (2014–2015), and 10 (2018–2019).

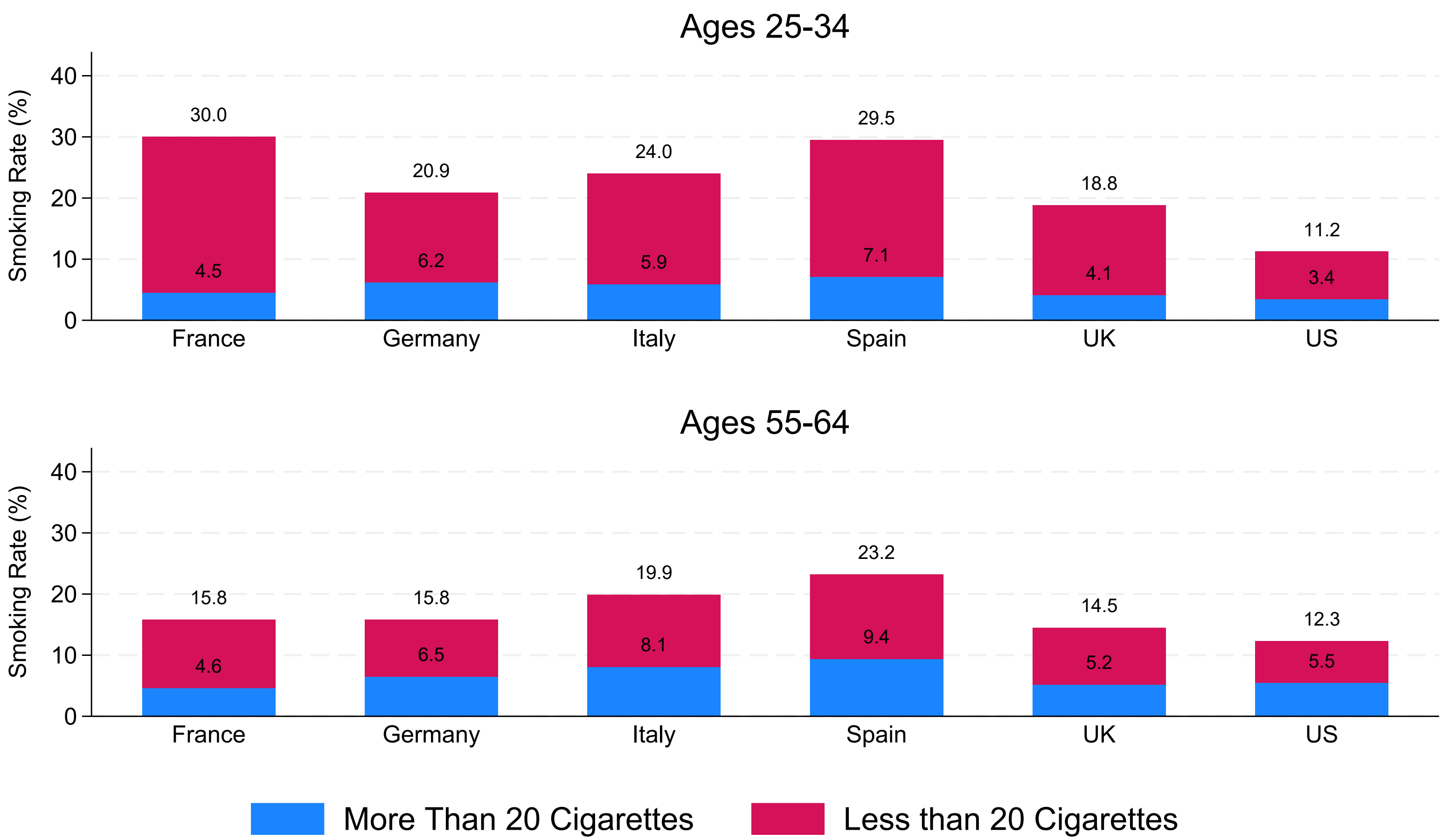

Figure A.10. Smoking Rates by Age Group for Six Selected Countries in 2014. *Notes:* The number at the top of each bar is the total daily smoking rate for the given age group and country. This number is equal to the combined height of the red and blue sections of the bar. The lower number corresponds to the population share of persons who smoke more than 20 cigarettes per day. This number corresponds to the blue section of the bar. *Sources:* European data are from Eurostat. US data are from the Tobacco-Use Supplement of the Current Population Survey.

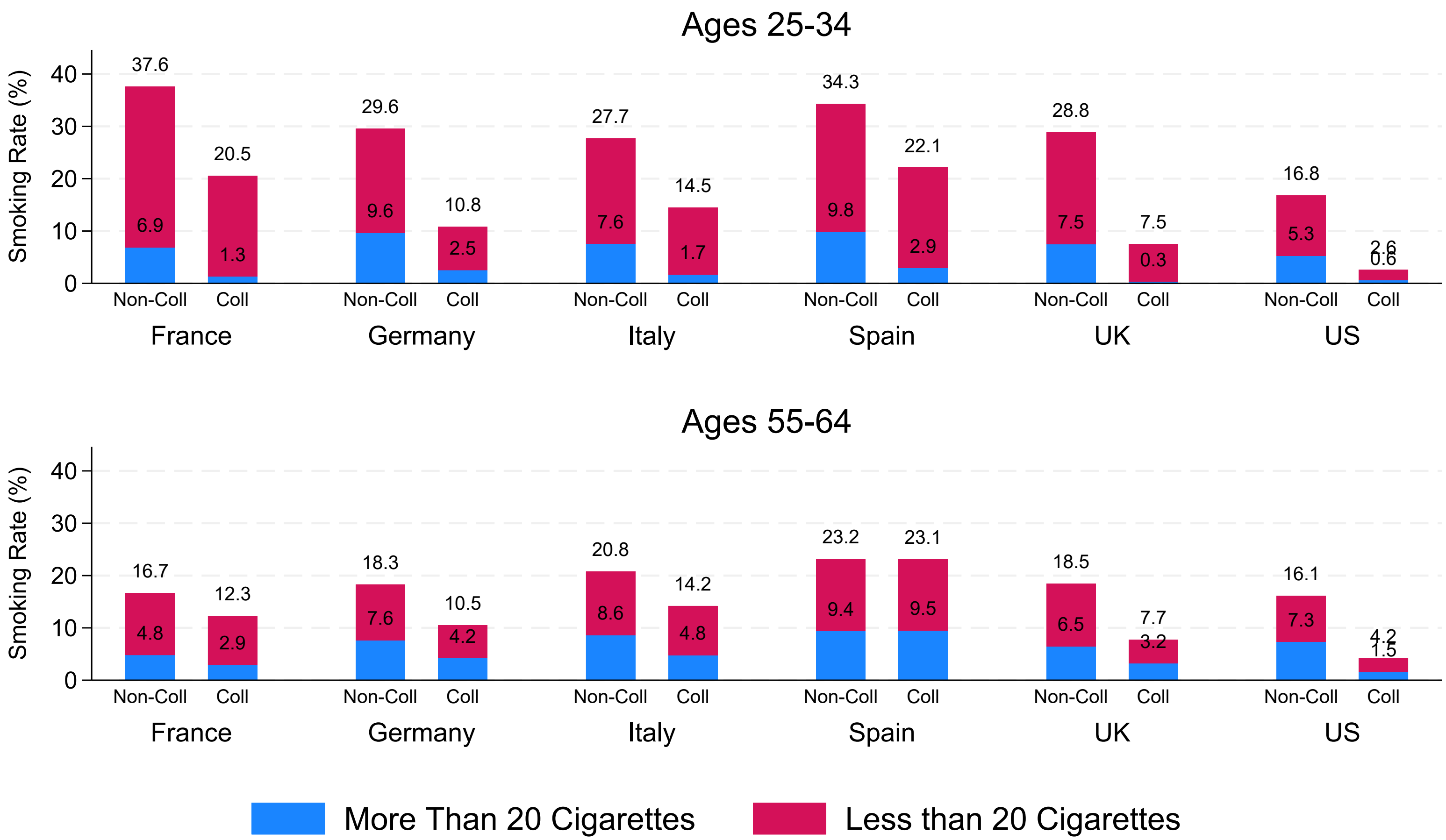

Figure A.11. Smoking Rates by Age Group and Education Level for Six Selected Countries in 2014. *Notes:* The number at the top of each bar is the total daily smoking rate for the given education level, age group and country. This number is equal to the combined height of the red and blue sections of the bar. The lower number corresponds to the population share of persons who smoke more than 20 cigarettes per day. This number corresponds to the blue section of the bar. For the European countries, the non-college rate is the population-weighted average of smoking rates for persons with education corresponding to ISCED11 categories 0–2 (lower secondary education and less) and categories 3–4 (upper secondary and post-secondary non-tertiary education). The college category corresponds to ISCED11 categories 5–8 (tertiary education). *Sources:* European data are from Eurostat. US data are from the Tobacco-Use Supplement of the Current Population Survey.

B Appendix: Data

B.1 County-level Data

For any county-level variable, we require a set of consistent county definitions over time so that our observations correspond to the same geographic area throughout the panel. Fortunately, county definitions are more consistent than any other sub-state geographic unit over time. However, some changes do occur. Furthermore, different data systems use different conventions for independent cities or other local areas that are similar to counties. Therefore, we must adopt the conventions of a given system and convert between the conventions of other systems to account for this.

A small number of counties combined or split during the sample period. To account for this, we use the largest definition of a given county or set of counties throughout the whole period. For example, in 1983, Yuma County (FIPS 4027) splits into a smaller Yuma County and La Paz County (FIPS 4012). For years in which the two are split in the data, we combine them and assign them the FIPS code of the original (4027). The following counties are edited across all data sources:[32]

- Miami-Dade county is recoded from 12025 to 12086 in all periods due to a renaming in 1997.
- Counties with FIPS codes 36005, 36047, 36061, 36081, and 36085 (which are the 5 FIPS codes associated wit New York City) are combined in all periods.
- Yuma County (FIPS 4027) and La Paz County (FIPS 4012) are combined in all periods.
- Counties 8001, 8013, 8014, 8059, and 8123 (alternately known as 8911, 8912, 8913, 8914, and 8014) are combined for all periods.
- Counties 30113, 30067, and 30031 are combined for all periods.
- Counties 35006 and 35061 are combined in all periods.
- Records with FIPS 13999 are dropped from mortality data, because this special county designation was given to all HIV deaths in Georgia in some years.
- County 46113 is recoded to 46102 in all periods.
- Counties 55078 and 55115 are combined into 55901 in all periods.

The state of Virginia has counties and independent cities that are considered county-equivalents by the Census. Some data sources, like SEER and the NCHS micro-level mortality data, do not combine these regions. Others, like the BEA, combine these into slightly larger regions. We aggregate to the higher level, using BEA definitions. However, in rare cases we must further aggregate these regions for intertemporal consistency. Furthermore, in early SEER data (before the early 80s), some county/city combinations in Virginia had different codes

[32]Thanks to David Dorn for compiling a list of county changes that was helpful in developing these definitions. See the list here: https://www.ddorn.net/data/FIPS_County_Code_Changes.pdf.

than the BEA data but were nonetheless combined. These codes need to be reassigned to the BEA definitions. We use a crosswalk to convert between the two definitions where necessary.[33]

After these combinations are made, the final panel contains 3063 counties with data from 1969 to 2021.

B.2 Mortality Rates

In general, to calculate mortality rates in a given county and year, we need to obtain mortality counts and population in that county and year. Before 1989, we obtain data on both population and mortality from the CDC's WONDER database. From 1989 onward, we use Restricted-Use Vital Statistics data from the National Vital Statistics System of the National Center for Health Statistics for mortality data. We gather population data from the Surveillance, Epidemiology, and End Results program of the National Cancer Institute for this period. In both systems, we drop data from Alaska, DC, and Hawaii for data reliability concerns.

B.2.1 Pre-1989 Mortality Rates

The WONDER database contains both deaths and population by county, age group, and cause from 1968 to 1988. We pull data by county for 25–34, 35–44, 45–54, and 55–64 age groups (which we use as age bins throughout), separately for all-cause mortality and for a set of specific causes. Some records are regarded as unreliable due to a small number of deaths occurring. We choose to not treat these differently than other records. After standardizing county definitions, we calculate the deaths and populations by year and age group for our counties, separately by cause. We then calculate the mortality rate per 100,000 people by county, year, and age group for each cause.

B.2.2 1989-2021 Mortality Rates

The NCHS Restricted-Use Data provides us with individual death records that contain, among other things, county of residence, county of death, demographics such as age and race, cause of death, and educational attainment, from 1989 to 2021 for all recorded deaths during that period. We convert both county of residence and county of death into our county basis before we aggregate deaths to the county and 10-year age group level for all-cause mortality and for a select group of causes for each year. For population, we take data from SEER from 1989 to 2020. We convert SEER to our county basis, then aggregate to the county and age group level by year to merge with the deaths data.

Data on the 2021 population is not available from SEER. To address this, we use Census population data from 2020 and 2021. We aggregate these data to our county definitions in 5 year age bins. We convert to county definitions for Connecticut, which changed its statistical areas in the 2020 Census from counties to planning regions. We then merge both years of data with SEER data from 2020. We estimate a linear regression of SEER population in

[33] A blog post on the topic and the crosswalk itself can be downloaded here: https://www.economy.com/support/blog/buffet.aspx?did=869A03D1-5D74-4376-A606-00A8C64DDB0B.

2020 on Census population in 2020, then predict SEER population in 2021 using that model and the Census population in 2021. This gives us an estimate of what SEER population would be in 2021 data. We then aggregate these population estimates to our 10-year age bins.

We combine our SEER population data from 1989 to 2020 with our population estimates from 2021 then merge on our death counts from the microlevel mortality data. Some counties have 0 deaths with nonzero population for certain age groups, which is plausible for small populations. After aggregating to 10-year age bins, no counties have a population of 0 with non-zero deaths, suggesting our population estimates are reliable. We then calculate the death rate per 100,000 by cause and age-group for each county and year.

Finally, we append the mortality rates that are calculated by county, year, and 10-year agegroup from the WONDER database with those from the microlevel mortality data.

B.2.3 Age-standardizing Mortality Rates

To adjust for the fact that older age groups tend to have higher mortality rates, which would distort relative mortality rates for counties with different age distributions, we age-standardize mortality rates using the 25–34, 35–44, 45–54, and 55–64 age bins and the 2000 Census standard age distribution. Specifically, we calculate

$$M_{cst} = \sum_{a=1}^{4} \frac{P_{a,2000}}{P_{2000}} M_{acst}$$

where M_{acst} is the age-group-specific mortality rate in year t, county c, state s, and age group a, $P_{a,2000}$ is the population in age group a in 2000, and P_{2000} is the total population in 2000.[34] Therefore, we now have age-standardized midlife mortality rates by cause for select causes by county and year.

B.2.4 Causes of Death

We combine multiple ICD-8, ICD-9 and ICD-10 codes into our own cause-of-death categories that are mostly consistent across years. For the most part, we are interested in all-cause mortality. Table B.4 details our categories and the corresponding ICD codes for the three systems.

We choose these categories and the corresponding ICD codes to represent major causes of death which are mutually exclusive.[35].

To properly convert between ICD-9 and ICD-10 mortality counts, we multiply death counts in each category by comparability ratios provided by the CDC, which inflate/deflate death counts by major cause groups so that they are consistent through time.[36]

[34] See this CDC statistical note for details: https://www.cdc.gov/nchs/data/statnt/statnt20.pdf

[35] For deaths related to alcohol and drugs, as well as suicide deaths, we use the list of ICD codes used in this Senate Joint Economic Committee report: https://www.jec.senate.gov/public/index.cfm/republicans/methodological-appendix-to-long-term-trends-in-deaths-of-despair/

[36] See this CDC page for more details on ICD comparability: https://www.cdc.gov/nchs/nvss/mortality/comparability_icd.htm

We calculate the mortality rate that is not otherwise specified by our select causes (by subtracting deaths of each cause by the total), and, where this is less than zero due to rounding and the use of comparability ratios, we set it equal to 0. This minor adjustment affects very few counties.

B.3 Mortality Rates by Education

Educational attainment is provided for some records in all years, but is frequently missing in 1989, 1990, and 1991. Two different definitions for educational attainment are used depending on the data year, one from 1989 and one from 2003. We map both onto our own education variable that records education into six categories: less than high school, some high school, high school graduate, some college, bachelor's degree, and advanced degree. We typically compare those with a college degree to those without.

Most of the processing associated with these mortality rates is conceptually the same as the combined mortality rates described above. We calculate the number of deaths by county, year, age group, and education then we calculate the relevant populations for this group. We then take the ratio of these to get the mortality rate, which we age-adjust to be at the county-year level.

However, there are some key differences. First, due to the lack of consistently available data on population by education, county, and year, we use data from the Census and ACS as described in the text. At the state level, we can calculate the relevant populations using the Current Population Survey (CPS) yearly from 1992 through 2021.

Second, some records are missing information on education. In the 1990s, this happens somewhat frequently, but becomes less common in later years. If we simply assigned deaths to the category of education which is recorded in the death record, we would undercount total deaths (and mortality rates) in a county-year by a potentially large amount.

B.3.1 State and County Populations by Education

To calculate both state and county populations by education (and age group), we calculate shares of the population by educational attainment, and multiply those shares by SEER population counts to get estimates of the respective populations.

At the county level, we use data from the decennial Census (for 1990 and 2000) and from the ACS (for 2010 onwards) that records population counts in a PUMA by age group, sex, race, and educational attainment for the 25–34, 35-44, 45-54, and 55-64 age groups. We compute the population of each educational group within each PUMA cell. Then, we convert to a county-basis using a geographic crosswalk from the Missouri Census Data Center Geocorr Applications. Each PUMA is mapped to one or more counties, and its population is divided into those counties with an allocation factor. The allocation factor is the share of the PUMAs total population that lies within the county. We multiply the PUMA-cell populations by this allocation factor then compute the total population within a cell at the county level.

When a county is only within one PUMA, this county-cell population is simply the PUMA-cell population multiplied by the allocation factor. When a county contains multiple PUMAs, this county-cell population is the sum of all the PUMA-cell populations that it contains. Finally, we compute the share of the population by educational attainment within each county-cell to multiply by the total population computed from SEER. By doing this, our education-specific populations match the totals in the overall populations. In performing these calculations, we make two key assumptions: 1) The population distribution across counties within a PUMA is the same for all demographic cells and 2) The education distribution is the same within the same demographic cells in all counties within a PUMA.

For population estimates in 1992, we compute population estimates in 1990 and 2000, then estimate the 1992 population within each county-education-cell as 0.8 times the 1990 estimate plus 0.2 times the 2000 estimate (equivalent to a linear interpolation across that 10-year period).

At the state level, we use data from the Current Population Survey to estimate the share of the population in each state with a college degree for each of our original 10-year age groups, by year from 1992 to 2021. We use the monthly CPS and aggregate to a year level. Because the CPS faces considerable measurement error and year-to-year noise, we smooth our estimates using a moving average with 1 lag and 1 lead.[37] We then merge these estimated shares with SEER data to calculate population counts.

B.3.2 Adjusting for Missing Education

We must also adjust the way we calculate death counts by education. Many records are missing education data, so we take a similar approach as in the population data. We calculate the share of deaths in each county and age-race-sex cell by educational attainment (using records that are not missing education), then multiply it by the total count of deaths within the county-year-age group cell. This methodology implicitly assumes that the records with missing education data have the same education distribution those that have education data.

We then merge these estimated death counts with the population data (after aggregating to the state level, for state-level data), calculate age-group specific mortality rates by geographic unit and year, and age-adjust with the 2000 Census age distribution.

Some counties are missing education for many records, especially in 1992. This causes concern over the reliability of estimates associated with these counties. Therefore, if a county is missing education data in more than half of all records within an age group for a given year, we record mortality rates as missing for all education categories. When aggregating death counts to the state or national level, we do not exclude the counts from such counties. However, a small number of county-demographic cells are missing education data for all entries. Therefore, after aggregating death counts to the state and nation level, the total number of deaths across education categories is below that computed without regards to education by a small amount. Given the small number of records this affects, it will have a negligible effect on our analysis.

[37]On the boundaries of our time series, we use just the current value and the lag, or the current value and the lead, depending on which boundary.

B.4 Per Capita Income

Besides mortality rates, we also extensively use per capita income at both the state and county level. We take both of these from the Bureau of Economic Analysis. Because we use the BEA's broader definition of counties, less processing must be done to harmonize the county definitions with our conventions. When we combine two or more counties together, we take the average of their per capita incomes, weighted by population.

B.5 Behavioral and Health Outcomes from the Behavioral Risk Factor Surveillance System (BRFSS)

We compute rates of smoking, obesity, and inactivity using survey data from BRFSS. We create indicator variables using the following definitions.

- Smoking: The respondent must have smoked at least 100 cigarettes in their lifetime, consider themselves an active smoker, and smoke regularly. This is the basis of a calculated variable in the BRFSS data.
- Obesity: The respondent has a body mass index over 30.
- Inactivity: The respondent has not exercised or performed recreational activities outside of work in the last 30 days.

We then compute means of each indicator variable within counties (both overall and conditional on education) using provided survey weights, across all years in which county is identified in the data, 1996 to 2010. We set a cell's given rate to missing if fewer than 50 records are used to compute that rate across the 15 year period.

At the state level, we are able to compute rates that change over time. We compute weighted means of each indicator variable by state and year, overall and across educational attainment, from 1992 to 2021.

Cause category	**ICD-8 Codes**	**ICD-9 Codes**	**ICD-10** Codes
Circulatory	390-398, 402, 404, 410-413, 420-429, 440, 441, 450	390-398, 402, 404, 410/425.4, 425.6/429.9	I00/I09.9, I11, I13, I20/I42.5, I42.7/I51.9
Cancer	140/162.0	140/162.0, 163/208	C00/C33.9, C35/C97
Lung cancer	162.1	162.2/162.9	C34
Lower Resp.	490-493, 518, 519.3	490/494	J40/J47
External	E800/E849, E861/E949, E960/E979, E980.4/E999	E800/E849, E861/E949, E960/E979, E980.6/E999	V01/X39, X46/X59, X85/Y09, Y16/Y86, Y88/Y89
Cerebrovascular	430/434, 436/438	430/434, 436/438	I60/I69
Diabetes	250	250	E10/E14
Alcohol	291, 303, 571.0, E860	291, 303, 305.0, 357.5, 425.5, 535.3 571.0/571.3, 790.3, E860	E24.4 F10, G31.2, G62.1, G72.1, I42.6, K29.2, K70, K85.2, K86.0, O35.4, P04.3, Q86, R78.0, X45, Y15
Suicide	E950/E959	E950/E959	X60/X84, Y87
Drugs	304, E850/E859, E980.0/E980.3	292, 304, 305.2/305.9, E850/E858, E980.0/E980.5	F11/F16, X40/X44, Y10/Y14

Table B.4. ICD Codes by Causes of Death.